A Distantly Supervised Method for Extracting Spatio-Temporal Information from Text

Seyed Iman Mirrezaei
University of Illinois at Chicago
Department of Computer Science
851 S. Morgan, Chicago, IL, USA
smirre2@uic.edu

Bruno Martins
IST and INESC-ID
University of Lisbon
Rua Alves Redol, 9, 1000-025 Lisboa, Portugal
bruno.g.martins@ist.utl.pt

Isabel F. Cruz
University of Illinois at Chicago
Department of Computer Science
851 S. Morgan, Chicago, IL, USA
isabelcfcruz@gmail.com

ABSTRACT

This paper describes TRIPLEX-ST, a novel information extraction system for collecting spatio-temporal information from textual resources. TRIPLEX-ST is based on a distantly supervised approach, which leverages rich linguistic annotations together with information in existing knowledge bases. In particular, we leverage triples associated with temporal and/or spatial contexts, e.g., as available from the YAGO knowledge base, so as to infer templates that capture new facts from previously unseen sentences.

CCS Concepts

•Information systems → Information extraction; •Computing methodologies → Information extraction;

Keywords

Spatio-temporal information; geographic information retrieval; distant supervision; text mining; open relation extraction

1. INTRODUCTION

Large cross-domain Knowledge Bases (KBs) such as YAGO [10], DBpedia [1], and Wikidata [15] have been constructed by Web companies and research communities to support search and question answering systems. These KBs were mostly built by large scale harvesting of facts from Wikipedia infoboxes, containing rich information about entities and about their relations. The information in the KBs consists of *facts*, which are triples following the format `<subject; relation; object>`.

Existing KBs contain a wide variety of spatial and temporal facts, although these are mostly *static*, in the sense that their truth value does not change over time and/or across geospatial regions. These static facts may include information about entities and their spatio-temporal properties. Examples include dates of major events or spatial information associated with people or companies (e.g., Facebook was founded on February 4, 2004; Barak Obama was born in Honolulu, Hawaii). Such information is useful for supporting innovative information retrieval services.

In addition to static facts, other facts may change as time passes and/or the location changes, and are therefore *dynamic*. For instance, information about countries may change over time, new products or movies are often released at different times in different countries, and soccer players get transferred between clubs. Associating dynamic facts in KBs with their spatio and/or temporal contexts is of great importance since dynamic facts are only valid inside those contexts. We define *temporal context* as a temporal expression (instant or time interval) associated to dynamic facts to show when they are valid. Conversely, a *spatial context* is a geospatial region where a dynamic fact is valid. Ideally, dynamic facts should be associated with an instant or time interval, corresponding to a *temporal context* (Bonn was the de facto capital of Germany *between 1949 and 1990*; Barack Obama is the president of the United States *from 2009 to 2016)*, and/or with a geospatial region where they are valid, i.e., their *spatial context* (Nintendo sold the color TV-game consoles only in *Japan*).

To access static and dynamic facts from a variety of web documents, information extraction methods can be used. Open-domain Information Extraction (OIE) systems often rely on distant supervision, using existing information (e.g., triples in KBs, together with matching sentences), which may be noisy, as examples to infer patterns. While state-of-the-art OIE systems address to some degree the extraction of static spatio-temporal facts, they have not properly addressed the extraction of spatial or temporal contexts. Some systems assign temporal contexts based on the time of document creation or publication [5, 7]. However, this assumption is not necessarily true for all facts within a document. For instance, the Wikipedia page that contains the *List of Presidents of the United States* includes facts whose temporal context is not related to the time of the document creation. Some previous studies have instead focused exclusively on the extraction of temporal and/or spatial information from text [14, 16]. However, the set of temporal and spatial relations that are considered in these studies is closed and is also relatively small in comparison to the types of relations that can be extracted with OIE systems.

This paper describes TRIPLEX-ST, an evolution of our TRIPLEX system [12], which focuses on the extraction of dynamic facts associated with temporal and/or spatial contexts from text. Using TRIPLEX-ST, sentences that express triples are first discovered using information from Wikipedia text, leveraging a bootstrapping method [9, 11]. Then, TRIPLEX-ST uses rich linguistic annotations (e.g., dependency relations, named entities, and lexical constraints) together with information available in existing KBs, specifically YAGO, to infer templates. Templates show how dynamic facts are associated to spatio-temporal contexts, or how static facts, can be expressed in textual resources. Finally, templates can capture new

SIGSPATIAL'16, October 31-November 03, 2016, Burlingame, CA, USA
© 2016 ACM. ISBN 978-1-4503-4589-7/16/10. . . $15.00
DOI: http://dx.doi.org/10.1145/2996913.2996967

facts from previously unseen sentences. The main contributions of this paper are as follows:

- We advance a novel method that focuses on the extraction of dynamic facts associated with spatio-temporal contexts from textual resources;
- We discuss how the proposed method can also address the extraction of static facts involving spatio-temporal information from textual resources, improving on our previous TRIPLEX system in these particular cases.

2. TRIPLEX-ST

Existing OIE systems extract triples from input sentences. A triple involves a subject and an object, which are typically noun phrases, and a relation phrase, which is a text fragment expressing a semantic relation (i.e., a predicate or property) between the subject and the object. The relations in triples can be expressed either by verb phrases (verb-mediated) or by noun phrases (noun-mediated).

TRIPLEX, and also the extension named TRIPLEX-ST, involves an offline stage of collecting training instances (i.e., sentences that match known triples), followed by the inference of extraction templates from these instances. The templates can then be used to extract new triples from text, and these triples are finally validated by a classifier. Our previously proposed TRIPLEX OIE system [12] only focused on noun-mediated relations related to the usage of nouns, adjectives, and appositions, but we now extend the system to also consider verb-mediated relations, envisioning the application to the extraction of spatio-temporal information from text.

TRIPLEX-ST extracts spatio-temporal information involving dynamic or static information about entities and their properties. It therefore extends the general model of triples by considering information regarding the temporal and/or spatial context that qualifies the facts expressed in triples, in the case of relations that involve dynamic information and if this information is available in the text. Triples should ideally be assigned to an instant or a time span, and/or to the geospatial region when and where they are valid. For instance, the triple `<Clinton; served as; President;>` `(Spatial context: United States; Temporal context: [1993 : 2001])` should be extracted from the sentence *In the United States, Clinton served as President, from 1993 to 2001.* Moreover, our TRIPLEX-ST system can also better identify whether there are semantic relations between spatio-temporal expressions and different types of named entities in textual resources, at least in comparison to the original TRIPLEX system. The subject type or the object type of these triples is either *Location* or *Date*. These triples thus involve static spatio-temporal information about entities and their spatio-temporal properties (e.g., a person's birth date or place of residence, the population of a city or town, capitals of countries, etc.) as expressed over text.

TRIPLEX-ST uses Wikipedia infobox values as well as triples in YAGO [10], DBpedia [1], Freebase [2] and in Wikidata [15] during its bootstrapping process. We use a recent Wikipedia English dump to extract all Wikipedia pages. Then, we query the different KBs according to the Wikipedia page ID, to determine the type of the page. Afterwards, we classify Wikipedia pages under the following types: *Person*, *Organization*, *Location*, or *Unknown*.

TRIPLEX-ST uses the Stanford NLP toolkit[1] to parse and chunk sentences, extract dependency relations, label tokens with named entity (NE) and with part-of-speech (POS) information, and perform coreference resolution. Specifically, the Stanford dependency parser discovers the syntactic structure of input sentences, producing a directed graph whose vertices are words and whose edges are syntactic relations between words [4]. For example, a dependency relation `nsubj<had, Bhutan>` exits between the governor word `had` and the dependent word `Bhutan` in Figure 1. TRIPLEX-ST uses dependency paths during the bootstrapping process. A dependency path connects words of a sentence by using dependency relations. In Figure 1, the arrows show the dependency path between word `Bhutan` and the token `770,000`.

We also complement the results from the Stanford NLP toolkit with those from other tools (e.g., the HeidelTime temporal expression resolver [13], or the AIDA system for named entity disambiguation [8]), and through resources such as WordNet[2] and VerbNet.[3] The HeidelTime temporal expression tagger is used to complement the results from Stanford NER, also classifying (e.g., according to classes such as date, time, duration, or set) and normalizing the recognized temporal expressions. TRIPLEX-ST also uses WordNet and VerbNet to enrich the annotations provided by the NLP pipeline. WordNet is a lexical database that classifies English words and phrases into sets of synonyms called synsets. VerbNet is in turn a hierarchical verb lexicon, including syntactic and semantic information for English verbs. The NLP pipeline also constructs several synsets for each Wikipedia concept. We rely on the different mentions (e.g., redirects and alternative names) that are associated to Wikipedia URLs and also in the hypertext anchors that point to a Wikipedia page. For example, the word *UN* is extensively used in Wikipedia to refer to the concept *United Nations*.

2.1 Bootstrapping Set Creation

We follow ideas from the OLLIE system [11] to build automatically two bootstrapping sets. These sets include sentences that are extracted from Wikipedia text, expressing dynamic facts associated with spatio-temporal contexts, or static facts.

First, we use YAGO triples associated with a spatial or a temporal context, together with a large set of Wikipedia pages, to create a bootstrapping set that expresses dynamic facts associated with spatio-temporal contexts. Then, we process the Wikipedia pages and their relevant infoboxes to construct the bootstrapping set for the extraction of static spatio-temporal facts.

In dynamic facts, we have that the subject can appear with multiple objects, or the object can appear with multiple subjects, depending on the spatial and/or temporal context. In the case of dynamic facts that are associated with temporal contexts, we found around 700,000 instances in the set of YAGO triples. YAGO has therefore a good coverage of dynamic facts associated with temporal contexts, although it is very poor in terms of spatial contexts. Still, YAGO includes information about the geographical location of several entities. We use the relation `<isLocatedIn>` in YAGO to find the spatial contexts of objects within dynamic facts. We found around 1,600 dynamic facts having a spatial context and we also have a total of 800 facts with both spatial and temporal contexts. These instances constitute the source of distant supervision to our methods. Afterwards, the sentence extractor matches phrases from the text of the Wikipedia pages with the corresponding dynamic facts and also with their spatio-temporal contexts. If there exist dependency paths connecting an object of a triple, the spatio-temporal context of the triple, and the synset of the Wikipedia page in a sentence, then the sentence is extracted and added to the bootstrapping set. For example, given the page for *Bhutan*, the extractor matches the triple associated to the temporal

[1] http://nlp.stanford.edu/software/corenlp.shtml

[2] https://wordnet.princeton.edu/

[3] https://verbs.colorado.edu/ mpalmer/projects/verbnet.html

context <Bhutan; population; 770,000;> (Temporal context: [2015]) with the sentence *Bhutan had a population of 770,000 People in 2015*. Notice that there exists a dependency path connecting the object token 770,000, the temporal context 2015, and the synset word Bhutan in Figure 1. Thus, the sentence *Bhutan had a population of 770,000 People in 2015* could be added to the bootstrapping set of dynamic facts.

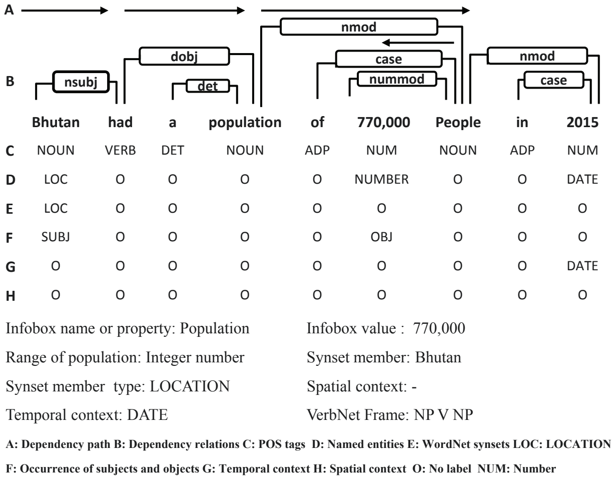
In the case of the bootstrapping set for static spatio-temporal facts, we process the extracted Wikipedia pages if their type is *Location*, or if the type of their infobox values is either *Location* or *Date*. A sentence extractor matches infobox values with phrases from the text of the corresponding Wikipedia page to create a bootstrapping set automatically. If there exists a dependency path between an infobox value and the synset of the Wikipedia page in a sentence, then the sentence is extracted and added to the bootstrapping set. For example, given the page for *Flacq District*, the extractor matches the infobox value 297.9 km^2 with the sentence *Flacq District has an area of 297.9* km^2. Since, there exists a dependency path between the infobox value 297.9 km^2 and the synset Flacq District, the sentence is added to the bootstrapping set of static facts.

In both cases, we also apply a constraint on the length of the dependency path between a synset member and an infobox value (an object) to reduce bootstrapping errors. This constraint sets the maximum length of the dependency path to 8, which was determined experimentally by examining the quality of both bootstrapping sets. Moreover, we use two methods, proposed by Intxaurrondo et al. [9], to remove noisy sentences automatically from a bootstrapping set. The first method computes the Triple-PMI between the entities (i.e., the subject and the object) of a triple and its corresponding labels, as extracted from a sentence [3]. A Triple-PMI score indicates a noisy sentence if it is below the threshold of 10^{-3}, which was determined experimentally. Intxaurrondo et al. [9] also proposed a method to compute the centroid of all labels for each property in a bootstrapping set. They believe that the noisy labels of a property are far from the centroid of the cluster of labels for that property. We keep 85% of the most similar labels to the centroid for each property, and discard the rest of the labels.

2.2 Template Extraction

Having the bootstrapping sets, the next step is to generate extraction templates from dependency paths of sentences in both bootstrapping sets. These templates show how the spatio-temporal contexts of triples, or how spatio-temporal facts, can be expressed in textual resources. When considering dynamic facts, the templates include the shortest dependency paths that connect the synset member, the object, and also the phrases that correspond to the spatial and/or temporal contexts of the triple. When considering static spatio-temporal facts, the templates include the shortest dependency path between a synset member (a subject) and an infobox value (an object). These paths are annotated by POS tags, named entity types, and WordNet synsets. VerbNet is also used to add syntactic frames and semantic restrictions for verbs in the dependency path of the template. Each template includes dependency relations, POS tags, named entity annotations, WordNet synsets, subject types, object types, and syntactic/semantic restriction of verbs in dependency paths. Figure 1 shows a template that can be used to extract dynamic facts associated to a temporal context.

Synset members of the Wikipedia page name may occur before or after infobox values in sentences. If there exists a dependency path between these values, independently of their position, the relevant template is extracted. For example, the synset member may also occur before the infobox value, as shown in the sentence *Bhutan had a population of 770,000 People in 2015*. In this case, the word *Bhutan* is the synset member and the token *770,000* is the infobox value (see Figure 1).



	Bhutan	had	a	population	of	770,000	People	in	2015
C	NOUN	VERB	DET	NOUN	ADP	NUM	NOUN	ADP	NUM
D	LOC	O	O	O	O	NUMBER	O	O	DATE
E	LOC	O	O	O	O	O	O	O	O
F	SUBJ	O	O	O	O	OBJ	O	O	O
G	O	O	O	O	O	O	O	O	DATE
H	O	O	O	O	O	O	O	O	O

Infobox name or property: Population — Infobox value : 770,000
Range of population: Integer number — Synset member: Bhutan
Synset member type: LOCATION — Spatial context: -
Temporal context: DATE — VerbNet Frame: NP V NP

A: Dependency path B: Dependency relations C: POS tags D: Named entities E: WordNet synsets LOC: LOCATION
F: Occurrence of subjects and objects G: Temporal context H: Spatial context O: No label NUM: Number

Figure 1: An example template resulting from a sentence annotated by the NLP toolkit.

2.3 Template Matching

Templates that involve a spatio-temporal context are used first during the step of template matching. The idea is to extract the context information, if possible. Afterwards, the templates for the extraction of static spatio-temporal facts are used. When extracting information from a new document, the NLP pipeline is first used to annotate the sentence. Then, we match the dependency parse of the sentence with the dependency paths of templates to identify the candidate subjects and objects. Afterwards, infobox names (properties) of templates are assigned to a candidate subject and a candidate object, derived from matching templates with subject types, object types, dependency relations, WordNet synsets, POS tags, syntactic frames and semantic restrictions of verbs in dependency paths, temporal context types, spatial context types, and named entity annotations. If the pipeline recognizes a temporal context and/or a spatial context, and if its type matches with the type of the context in the template, then the recognized context is also attached to the triple. For example, the candidate subject Iran and the candidate object 65 million are recognized after matching the dependency path of the sentence *Iran had a population of 65 million people in October 2010* with the template in Figure 1. The spatio-temporal expressions in the sentence are recognized by the NLP pipeline. Then, the dependency path between the candidate subject, the candidate object, and spatio-temporal expressions are annotated by TRIPLEX-ST. Since all annotations of the dependency path are matched with the template in Figure 1, the dynamic fact <Iran; had the population; 65 million> (Temporal context: [October : 2010]) is extracted from the sentence.

A classifier is finally used to validate the triples that match templates, taking inspiration on the confidence function from systems like OLLIE [11] and ReVerb [6]. We pre-trained logistic regression classifiers by using 1000 manually labeled facts extracted from Wikipedia pages, using one model for static relations and another model for dynamic relations. A confidence score is obtained from the probability computed by the classifier. The set of features used in those models are based on syntactic dependency relations, named entities, and lexical constraints.

3. EVALUATION

Due to the length constraints of a short paper, we omit here many of the details regarding the evaluation of TRIPLEX-ST, leaving their presentation to a future publication.

We mostly leveraged the automated evaluation approach proposed by Bronzi et al. [3] to evaluate TRIPLEX-ST. The automatic evaluation procedure uses existing knowledge bases and a Triple-PMI metric to verify whether a fact is correctly extracted or not. We also used a manual evaluation procedure where a human judge verifies each possible extracted fact. Still, manual evaluation was only made with a very small dataset.

The total number of triples having dynamic temporal and/or spatial contexts in YAGO is 709,012. The total number of triples involving dynamic spatio-temporal contexts that are extracted by TRIPLEX-ST, from a set of 50,000 sentences, is of 84,134 from a maximum possible value of 400,140 triples considered in the automatic evaluation. Overall, with the automated evaluation procedure, we estimate an F1 score of 0.43 for the extraction of dynamic facts, and an F1 score of 0.45 for the extraction of static spatio-temporal facts. Similar results in terms of the F1 metric were obtained with the manual evaluation procedure.

4. CONCLUSIONS

We presented TRIPLEX-ST, a novel system for the extraction of static spatio-temporal facts and of dynamic facts associated with spatial and/or temporal contexts. We use a distant supervision approach, which leverages rich linguistic annotations together with information available in Wikipedia and in other knowledge bases.

Due to the space constraints of a short paper, we are not describing here many of the details regarding TRIPLEX-ST's components. We are also not detailing the method for the removal of noise from the bootstrapping sets, or the evaluation procedure. We plan to present detailed experimental results in a later and more extensive publication, together with an extensive analysis on the extraction errors that were observed. The evaluation has, for the most part, leveraged an adapted version of the automated evaluation procedure advanced by Bronzi et al. [3], together will small-scale tests leveraging manual validation. Overall, with the automated evaluation procedure, we estimate an F1 score of 0.43 for the extraction of dynamic facts, and an F1 score of 0.45 for the extraction of static spatio-temporal facts. Our tests have also confirmed that TRIPLEX-ST could outperform previous systems (e.g., OLLIE) on the extraction of static facts, while at the same time extending OIE in the direction of considering dynamic information.

Acknowledgments

Work at UIC was partially supported by a grant from the Bloomberg Philanthropies and by NSF awards CNS-1646395, III-1618126, CCF-1331800, and III-1213013. Bruno Martins was supported by the Portuguese FCT, through the project grant with reference UID/CEC/50021/2013.

5. REFERENCES

[1] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. DBpedia: A Nucleus for a Web of Open Data. In *International Semantic Web Conference (ISWC)*, pages 722–735, 2007.

[2] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor. Freebase: A collaboratively created graph database for structuring human knowledge. In *ACM SIGMOD International Conference on Management of Data (SIGMOD)*, pages 1247–1250, 2008.

[3] M. Bronzi, Z. Guo, F. Mesquita, D. Barbosa, and P. Merialdo. Automatic Evaluation of Relation Extraction Systems on Large-scale. In *Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction (AKBC)*, pages 19–24, 2012.

[4] M.-C. De Marneffe and C. D. Manning. *Stanford Typed Dependencies Manual*. Stanford University, 2013.

[5] L. Derczynski and R. Gaizauskas. Information Retrieval for Temporal Bounding. In *ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR)*, pages 129–130, 2013.

[6] A. Fader, S. Soderland, and O. Etzioni. Identifying Relations for Open Information Extraction. In *Conference on Empirical Methods on Natural Language Processing (EMNLP)*, pages 1535–1545, 2011.

[7] G. Garrido, A. Peñas, B. Cabaleiro, and A. Rodrigo. Temporally Anchored Relation Extraction. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 107–116, 2012.

[8] J. Hoffart, M. A. Yosef, I. Bordino, H. Fürstenau, M. Pinkal, M. Spaniol, B. Taneva, S. Thater, and G. Weikum. Robust Disambiguation of Named Entities in Text. In *Conference on Empirical Methods on Natural Language Processing (EMNLP)*, pages 782–792, 2011.

[9] A. Intxaurrondo, M. Surdeanu, O. L. de Lacalle, and E. Agirre. Removing Noisy Mentions for Distant Supervision. In *Conference of the Spanish Society for Natural Language Processing*, pages 41–48, 2013.

[10] F. Mahdisoltani, J. Biega, and F. M. Suchanek. YAGO3: A Knowledge Base from Multilingual Wikipedias. In *Conference on Innovative Data Systems Research*, 2015.

[11] Mausam, M. Schmitz, R. Bart, S. Soderland, and O. Etzioni. Open Language Learning for Information Extraction. In *Conference on Empirical Methods on Natural Language Processing (EMNLP)*, pages 523–534, 2012.

[12] S. I. Mirrezaei, B. Martins, and I. F. Cruz. The Triplex Approach for Recognizing Semantic Relations from Noun Phrases, Appositions, and Adjectives. In *The Semantic Web: ESWC Satellite Events, Revised Selected Papers*, volume 9341, pages 230–243. 2015.

[13] J. Strötgen and M. Gertz. A baseline temporal tagger for all languages. In *Conference on Empirical Methods on Natural Language Processing (EMNLP)*, pages 541–547, 2015.

[14] A. Vempala and E. Blanco. Complementing Semantic Roles with Temporally Anchored Spatial Knowledge: Crowdsourced Annotations and Experiments. In *National Conference on Artificial Intelligence (AAAI)*, pages 2652–2658, 2016.

[15] D. Vrandečić and M. Krötzsch. Wikidata: A Free Collaborative Knowledge Base. *Communications of the ACM*, 57:78–85, 2014.

[16] J. O. Wallgrün, A. Klippel, and T. Baldwin. Building a Corpus of Spatial Relational Expressions Extracted from Web Documents. In *ACM SIGSPATIAL Workshop on Geographic Information Retrieval (GIR)*, pages 1–8, 2014.